# Szemantikuskeret-illesztés és az IE rendszer automatikus kiértékelése

Farkas Richárd[1], Konczer Kinga[2], Szarvas György[1]

[1] MTA SZTE Mesterséges Intelligencia Tankszéki Kutatócsoport
{rfarkas, szarvas}@inf.-szeged.hu
[2] Szegedi Tudományegyetem
kinga.konczer@hungary.org

**Kivonat:** Frametagger az SZTE Nyelvtechnológiai Csoportjának szemantikuskeret-illesztő programja, ami a gazdasági rövidhírek szereplőinek azonosítására született. A program az NKFP 2/017/2001 projekt[1] keretében, a Nyelvtudományi Intézet által elkészített, majd az SZTE által bővített keretekre és szemantikus táblázatokra épül. A program a szegedi IEToolChain[2] információkinyerő rendszer végső modulja. Előadásunkban bemutatjuk az IEToolChain kiértékelésére született Benchmark programot is, aminek célja, hogy pontos képet kapjunk arról, hogy az IEToolChain egyes moduljainak javítása, cseréje hogyan befolyásolja az egész rendszer hatékonyságát.

## 1 Szemantikuskeret-illesztés

Az információkinyerés célja a lényeges információ megjelölése és összegyűjtése dokumentumokból. A működő rendszerek általában megelégszenek a mondatok fontosabb szereplőinek azonosításával (az általános szemantikus szerepcímkézési feladattal [3] szemben, ahol a cél az összes ige vonzatkörnyezetének meghatározása) anélkül, hogy részletes szintaktikai ill. szemantikai elemzést végeznének.

Rendszerünkben a mondat szereplőinek azonításához a mondat ún. felszíni elemzését és egy szemantikuskeret-halmazt használunk fel. A keretek eseményeket írnak le azok szereplőinek szintaktikai és szemantikai megkötésein keresztül. Esetünkben tehát az információkinyerés a keretek célszavának illetve többi szerepének illesztése a mondatra.

### 1.1 Frametagger

Az Frametagger feladata, hogy az IEToolChain korábbi moduljai által előállított szintaktikailag elemzett szövegeken megtalálja és bejelölje a legjobban illeszkedő szemantikus szerepeket az előre definiált kerethalmaz alapján.

Frametagger inputját tehát a szintaktikailag (mondat- és szószegmentált, szófajilag egyértelműsített, NP taggelt) bejelölt szöveg, szemantikus táblázatok és a kerethalmaz alkotják. A – Nyelvtudományi Intézet által elkészített – szemantikus táblázatok 5471

főnévi és 3972 melléknévi jelenést osztályoznak (osztályok pl.: intézmény, absztrakt, cselekvőképes stb.)

Az általunk használt kerethalmaz a céginformációs gazdasági rövidhírek két témakörét írják le, a tulajdonosváltást és az intézménynyitást. A 71 darab keret szintaktikai és szemantikai megkötésekkel él az egyes szerepekre. A szükséges szemantikai információkat a szemantikus táblázatok alapján tölti ki a program.

Az NKFP 2/017/2001 projekt keretében elkészült kereteket az alábbiakkal bővítettük ki:

1. *Célszó* fogalmának bevezetése. Minden keretben a – korábbi szerepek közül – kijelöltünk pontosan egy célszót. A célszó általában ige (pl.: „megvásárol") de lehet más is, pl.: „alapkő". Egy illesztést csak abban az esetben tekintünk helyesnek ha a célszó illesztésre került, és a célszón felül legalább további egy szerep illeszkedik.
2. A célszavakon kívüli szerepekhez *prioritási* értéket vettünk fel. A szerep prioritási értéke megmutatja, hogy a szerep mennyire fontos az adott keretben a többi szerephez viszonyítva.
3. A szerepekhez különböző *pozícióbeli megkötések*et is adtunk. Azon felül, hogy keretmegszorítások közt megadható, hogy az egyik szerep a másik függvénye (azaz csak akkor illeszthető, ha a másik szerep illesztett), azt is meghatározhatjuk, hogy a függő szerep a függvényhez képest balra, jobbra helyezkedik-e el a mondatban, vagy közvetlen bal ill. jobb szomszédja-e. Erre elsősorban a birtokos illetve egyéb szerkezeteknél van szükség.

Mivel a mondat szavai, szerkezetei és a keretek szerepei egy (n*m-es) hozzárendelési feladatot határoznak meg, célszerű volt, hogy a programot az alábbi egyszerű algoritmus alapján építsük fel:

```
minden(mondatra) {
    minden(keretre) {
        költségmátrix kitöltése;
        magyar módszer végrehajtása;
    }
    legolcsóbb hozzárendelések bejelölése;
}
```

A hozzárendelési feladat kitöltése két részből tevődik össze, először minden (szó;szerep) párra megvizsgáljuk, hogy az adott megkötéseket teljesíti-e, majd a lehetséges illesztésekhez heurisztikaértéket számítunk. A felhasznált *heurisztikák* a következők: prioritási érték, tulajdonnév, mélység a szintaktikai fában.

Az olyan esetek tették szükségessé a mélységheurisztika hozzáadását, amikor a legfelső szintű szintaktikai egység több szerepből áll (pl: „28 százalékos részesedést"). A program szavakat feleltet meg a szerepeknek, de az illesztett szavak helyett azt a legmagasabb szintű nyelvtani szerkezetet jelöli be, amelyiknek az adott szó a feje.

A feladat magyar módszerrel történő megoldása időigényes, viszont az összes lehetséges megoldás által meghatározott térben keres, így nem veszíthetünk el megoldásokat.

### 1.2 Vizualizáció

A Frametagger outputja egy szemantikai információkkal bővített XML állomány, aminek átlátása a felhasználó számára igen komplikált. Ezért fejlesztettünk egy modult, ami az XML fájlt két felhasználóbarát formátumba konvertálja:

1. Egy HTML fájl generálódik, amelyben a megtalált szerepek különböző színekkel vannak jelölve, a szerep típusa pedig megjegyzésben jelenik meg. Ezen felül miden mondat után táblázatos formában is megjelennek a mondat különböző szerepei.
2. Egy Excel táblázatot is készítünk, amelyben egy munkalapon láthatjuk az azonos témájú híreket. A táblázat sorai egy-egy mondatot, oszlopai az egyes szerepeket tartalmazzák. Ennek segítségével könnyen készíthetünk komplex kimutatásokat (pl.: „Milyen cégeket vásárolt fel az OTP?")

## 2 Benchmark

Miután összeállt az egységes szegedi IEToolChain információkinyerő modullánc tudatában voltunk, hogy az egyes modulok külön-külön (tökéletes bejövő adatok mellett) milyen helyesen működnek, de nem tudtuk, hogyan befolyásolják a rendszert, mint egységet vizsgálva.

Egy olyan eszközt fejlesztettünk ennek vizsgálatára, ami egy etalonhoz hasonlítva nemcsak a végeredményről közöl (pontossági és találati) értékeket, hanem megpróbálja a helytelen (nem a legmegfelelőbb) illesztéseknél meghatározni, hogy mi a hiba oka és így melyik modul okolható érte.

Etalonnak a Szeged Korpusz NewsML részkorpuszából [4] 176 db hírt (285 mondatot) leválasztottunk. A – szintaktikailag már korábban annotált – mondatokat a kerethalmazhoz igazodva szemantikailag is bejelöltük. Ezt a mondathalmazt kivettük az összes tanuló algoritmust használó IEToolChain modul tréninghalmazából, így az tekinthető ismeretlen szövegnek.

A kiértékeléshez az alábbi hibakategóriákat határoztuk meg:

1. **Topikhiba**: ha az illesztett keret nem abba a témakörbe tartozik, mint a bejelölt keret.
2. **Feleslegesen felismert szerep**: olyan szerepek, amelyeket a gépi elemzés bejelölt, viszont az etalonbeli mondatban nem szerepelnek.
3. **Mondatszegmentálási hiba**: a program azért illesztette a szerepet helytelenül, mert az etalonbeli szereplőt külön mondatba szeparálta a mondatszegmentáló modul.
4. **POS hiba**: azért nem sikerült az illesztés, mert a helyes szerep MSD kódja nem egyezik meg a releváns helyeken a gépi elemző által adott kóddal.
5. **Lefedés**: azért sikertelen az illesztés, mert egy másik szerep eltakarja a felismerendő szavakat. Ez tulajdonképpen a fedő szerep hibája.
6. **NP hiba**: akkor tekintünk egy hibát NP hibának, ha a bejelölt illetve felismert szerepek közül az egyik a másik részhalmaza.

7. **Tagmondathiba**: a felismert szerep másik tagmondatba esik, mint a célszó. (az etalonban jelezve vannak a tagmondathatárok, viszont IEToolChainben nincs tagmondat-határolás)
8. **Igekötőhiba**: a gépi elemzés ugyan megtalálta az igét, de annak elváló igekötőjét nem jelölte be.
9. **Egyéb hiba**

A program az etalonbeli mondatokhoz hasonlítja egy TEI[5] kódolásnak megfelelő fájlhalmaz mondatait. Így a program megteremti a platformot arra is, hogy különböző magyar (gazdasági híreket feldolgozó) információkinyerő rendszereket, illetve azok moduljait (részfeladatokat végrehajtó egységeit) összehasonlíthassuk.

## 3 Eredmények és jövőbeni tervek

Az előző fejezetben bemutatott módszertan alapján a szegedi IEToolChain rendszer 70,2% pontossággal és 70,3% találati aránnyal működik. A két legjelentősebb hiba (és hibákon belüli arányuk) az NP hiba; 44% és a felesleges szerep; 29%. Mindössze 1 mondatnál követ el topikhibát a gépi elemzés (két témakör esetén).

Ha az illesztés jóságát másképp definiáljuk, és részleges egyezéseket (NP hibás illesztések tulajdonképpen a helyes szerepre találnak rá, csak nem ismerik fel azt pontosan) is elfogadjuk jó illesztésnek, akkor IEToolChain 83,4% F mértéket[1] produkál. Ezek alapján jogosan jelenthetjük ki, hogy a szegedi információkinyerő rendszer jelentős időt takaríthat meg – mint előfeldolgozó – egy manuális elemző számára.

Jelenleg folyamatban van a keretbeli -keretekben már szereplő- pozíciómegkötések Frametaggerbe történő beépítése, valamint az egyes részfeladatok alternatíváinak modulláncbeli tesztelése. Ezektől a javításoktól az IEToolChain további javulását várjuk.

A jövőben szeretnénk a Frametagger elé egy témaosztályozó modult beilleszteni. Ugyanis – mint az a 1.1 fejezetben látható – jelenleg a kerethalmazban nincs semmilyen különbség a két – jelenleg keretekkel lefedett – témabeli keretek közt. Azaz tulajdonképpen a megtalált keret azonosítja a témakört. A témák (elő)osztályozására feltétlenül szükség lesz, amikor a témakörök száma emelkedni fog.

Most végezzük ezen felül a teljes szintaxis felismerését végző modul integrálását az IEToolChainbe. Ez felveti a kérdést, hogy a kézzel kialakított Benchmark-hibaosztályok meddig és milyen áron bővíthetőek. Az elkövetkezendőkben szeretnénk megvizsgálni, hogy az általános, fatávolság alapú összehasonlítások versenyezhetnek-e a Benchmark specialitásokat kihasználó összehasonlításával.

[1] Az F mérték a pontossági és találati arány harmonikus közepe.

## Bibliográfia

1. Prószéky Gábor: Automatikus információszerzés gazdasági rövidhírekből. MSzNy 2003 (2003) 161–166
2. Alexin Zoltán, Gyimóthy Tibor, Csirik János: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004), beküldve, Szeged, Magyarország, (2004).
3. Xavier Carreras and Lluís Márques: Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. Proceedings of CoNLL-2004 (2004) 89–97
4. Csendes Dóra, Csirik János, and Gyimóthy Tibor: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In Sojka et al. [SKP04], pages 41–47.
5. Oravecz, Cs., Váradi, T.: TEI Encoding of the Hungarian Explanatory Manual Dictionary. In Kiefer et al. (eds.) Papers in Computational Lexicography COMPLEX'99, 1999, pp. 229-236